Scientific Methods for Health Sciences: Applied Inference (HS851): Fall 2014
http://www.socr.umich.edu/people/dinov/2014/Fall/HS851
Homework 3 Solutions

Use the SOCR Neuroimaging dataset of visceral pain, irritable bowel syndrome, ulcerative colitis, and Crohn's disease [1,2].

Problem 1: Investigate the effect of the grouping (Group), separating the 3 cohorts, on the gray matter brain volume (GMV), i.e., you can take group as a categorical variable and the GMV values as the observations. Use the http://www.socr.ucla.edu/htmls/ana/ANOVA1Way_Analysis.html SOCR 1-Way ANOVA applet, or R to carry the calculations. Provide all results and explicitly state your inference.

Step 1: Generate null and alternative hypotheses and choose a confidence level. We will choose an *a priori* false-positive rate $\alpha=0.05$.

H_0: $\mu_1-\mu_2=0$, where μ_i represents the population mean of group *i*. There is no difference in gray matter volume (GMV) among the 3 groups, i.e., there is no relationship between the presence of ulcerative colitis or IBS and GMV.
H_A: $\mu_1-\mu_2\neq0$. There is a relationship between the presence of ulcerative colitis and/or IBS and GMV.

Step 2: Visually explore the data.

Group 1=Control; Group 2=IBS; Group 3=Ulcerative Colitis
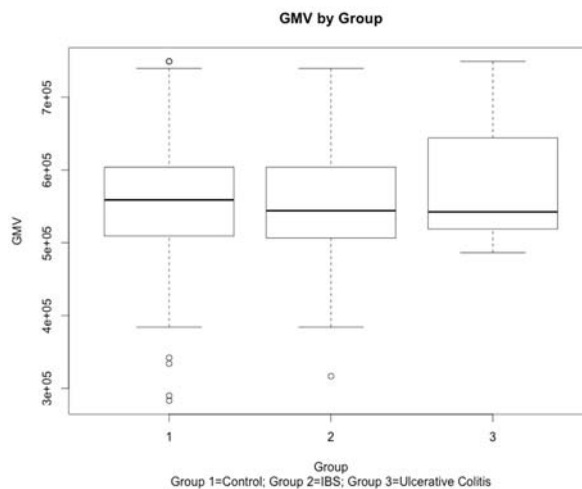
Step 3: Quantitative analysis
This side-by-side boxplot does not indicate much difference among the groups with respect to GMV. Does a quantitative analysis agree with this intuitive conclusion? We will use an ANOVA to find out.

```
> anova.gmv<-aov(GMV~Group, data=neuro.dat)
> summary(anova.gmv)
            Df    Sum Sq      Mean Sq      F value  Pr(>F)
Group       2     6.918e+09   3.459e+09    0.583    0.559
Residuals   332   1.968e+12   5.928e+09
```

Below are the parallel results from SOCR:

Sample Size = 335

---

[1] Note 1: The full R code is included as an appendix at the end.
[2] Note 2: The lines in the dataset corresponding to Group=5 were removed. There were only 2 individuals in this group.

Independent Variable = Group
Dependent Variable = GMV

Results of One-Way Analysis of Variance:
Standard 1-Way ANOVA Table. See:
http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_ANOVA_1Way
===

VarianceSource	DF	RSS	MSS	F-Statistics	P-value
TreatmentEffect (B/w Groups)	2	6917851866.852	3458925933.426	.583	0.558545154841116
Error	332	1968253851160.922	5928475455.304		
Total:	334	1975171703027.774			

===

Model:
Degrees of Freedom = 2
Residual Sum of Squares = 6917851866.852
Mean Square Error = 3458925933.426

Error:
Degrees of Freedom = 332
Residual Sum of Squares = 1968253851160.922
Mean Square Error = 5928475455.304

Corrected Total:
Degrees of Freedom = 334
Residual Sum of Squares = 1975171703027.774

F-Value = .583
P-Value = 0.558545154841116

R-Square = .004

Step 4: Interpret the results.
The quantitative analysis does agree with our visual assessment. We fail to reject the null hypothesis. There is not sufficient evidence to conclude that there is a difference in GMV among groups. Assuming the null hypothesis is true (i.e., the mean GMVs of the groups are equal), if we repeated this experiment many times, we would expect to see results this rare or rarer 55.9% of the time.

Problem 2: Use [2-way ANOVA](http://...) (using *gender* and *group* as the 2 factors) on the *R_insular_cortex* volume (insular shape has been reported to be associated with visceral pain, irritable bowel syndrome, and ulcerative colitis). Repeat the analysis using the counterlateral *L_insular_cortex* volume. Report your findings and interpret the results.

Step 1: Generate null and alternative hypotheses and choose a confidence level. We will choose $\alpha=0.05$.

H_01: (Factor 1: Group) There is no difference in right insular cortex volume among the 3 intestinal condition groups.
H_02: (Factor 2: Gender) There is no difference in right insular cortex volume between the sexes.
H_03: (Interaction) There is no interaction between the two factors (i.e., group and sex) that affects right insular cortex volume.

H_A1: (Factor 1: Group) There is a difference in right insular cortex volume among the 3 groups.
H_A2: (Factor 2: Gender) There is a difference in right insular cortex volume between the sexes.
H_A3: (Interaction) There is an interaction between group and sex that affects right insular cortex volume.

Step 2: Visually explore the data. Here, I provide a depiction of the mean of each group (where group is determined by the combination of sex and intestinal condition, i.e., there are 6 groups) along with a 95% confidence interval for each estimate (i.e., estimate $\pm$ 1.96 * SE(estimate)).
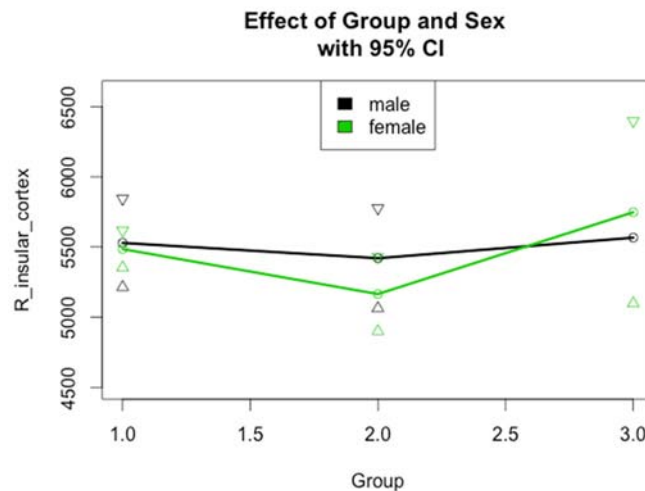
To consider the factor of sex, we look within each Group (i.e., compare the three pairs of points that lie above Groups 1, 2, and 3). We see that sex does not seem to have much effect because the points are close to each other, and the confidence intervals have a great deal of overlap.

Next we can examine the effect of Group by looking at the vertical distance between points of the same color. There seems to be a bit more difference here. Group 2 looks lower than both Groups 1 and 3; although, when we look at the confidence interval, it is not clear that this will be significant.

Finally, we see that the two lines are not parallel to each other, they diverge and then cross. This is indicative of the possibility of an interaction between sex and Group.

Step 3: Quantitative analysis:

```
> anova.r.ins<-aov(R_insular_cortex~Group*Sex, data=neuro.dat)
> summary(anova.r.ins)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	2	5691085	2845542	2.768	0.0643 .
Sex	1	680373	680373	0.662	0.4165
Group:Sex	2	631837	315918	0.307	0.7356
Residuals	329	338228335	1028050		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The quantitative analysis shows us that there are no significant effects for α=0.05. The main effect of group has the smallest p-value, at 0.06, but is not below our threshold of α=0.05. We therefore fail to reject the null hypothesis and conclude that these data do not provide evidence for a difference in the mean right insular cortex volume by sex, bowel disorder or an interaction of the two.

Here are the parallel results from SOCR (See below for the settings used in SOCR for left insular cortex):

Sample Size = 335
Dependent Variable = R_insular_cortex
Independent Variable(s) =   Group   Sex   Interaction Group:Sex

*** Two-Way Analysis of Variance Results ***

Standard 2-Way ANOVA Table. See:
http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_ANOVA_2Way
===

VarianceSource	DF	RSS	MSS	F-Statistics	P-value
MainEffect:Group		2	5691084.745	2845542.373	2.768 .064
MainEffect:Sex	1	352425.403	352425.403	.343	.559
Interaction Group:Sex		2	699037.743	349518.871	0.307	0.7356
Error	329	338228335.151	1028049.651		
Total:	334	345231629.182			

===

Variable: Group
Degrees of Freedom = 2
Residual Sum of Squares = 5691084.745
Mean Square Error = 2845542.373
F-Value = 2.768
P-Value = .064

Variable: Sex
Degrees of Freedom = 1
Residual Sum of Squares = 352425.403
Mean Square Error = 352425.403
F-Value = .343
P-Value = .559

Variable: Interaction Group:Sex
Degrees of Freedom = 2
Residual Sum of Squares = 699037.743
Mean Square Error = 349518.871
F-Value = 659.216
P-Value = .000

Residual: Degrees of Freedom = 329
Residual Sum of Squares = 338228335.151
Mean Square Error = 1028049.651
F-Value = 1.362
P-Value = 0.23805506448165337

R-Square = .017

We can repeat the procedure for the data from the left side (left insular cortex). The hypotheses are identical except that 'right' is replaced by 'left'.

Here is the parallel graphical representation of the data:

Effect of Group and Sex with 95% CI
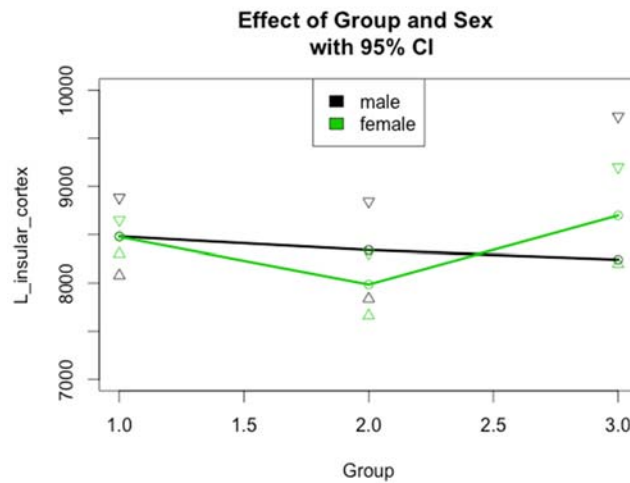
The data appear to be very similar to the right side. Perhaps there is slightly more difference among groups for females in this one. Also, we can note that for males, the pattern is slightly different; group 3 is lower than group 2 rather than higher. However, given the confidence intervals, we would not make much of this observation.

Here are the results of the ANOVA:
> anova.l.ins<-aov(L_insular_cortex~Group*Sex, data=neuro.dat)
> summary(anova.l.ins)

|	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	2	12389470	6194735	3.695	0.0259 *
Sex	1	806426	806426	0.481	0.4885
Group:Sex	2	1899498	949749	0.566	0.5681
Residuals	329	551643332	1676727		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This time, we find that the effect of Group is in fact significant at the chosen confidence level (i.e., $\alpha=0.05$). We therefore reject H_01 and conclude that there is enough evidence to support H_A1. The data suggest that intestinal condition is associated with left insular cortex volume. We note from the pattern in the plot above that group 2 (the IBS group) has the smallest left insular cortex volume. We fail to reject the other null hypotheses and conclude that there is not significant evidence to suggest that there is a relationship between sex and left insular cortex volume, nor does there appear to be a significant interaction between group and sex.

And here are the settings used in SOCR:

Some graphical output

Residual Normal QQ Plot

Scatter Plot of L_insular_cortex vs. Group
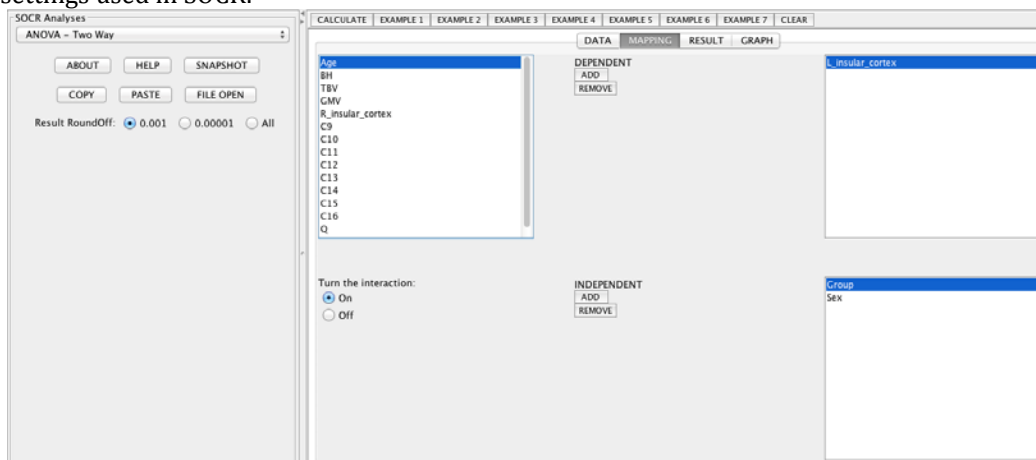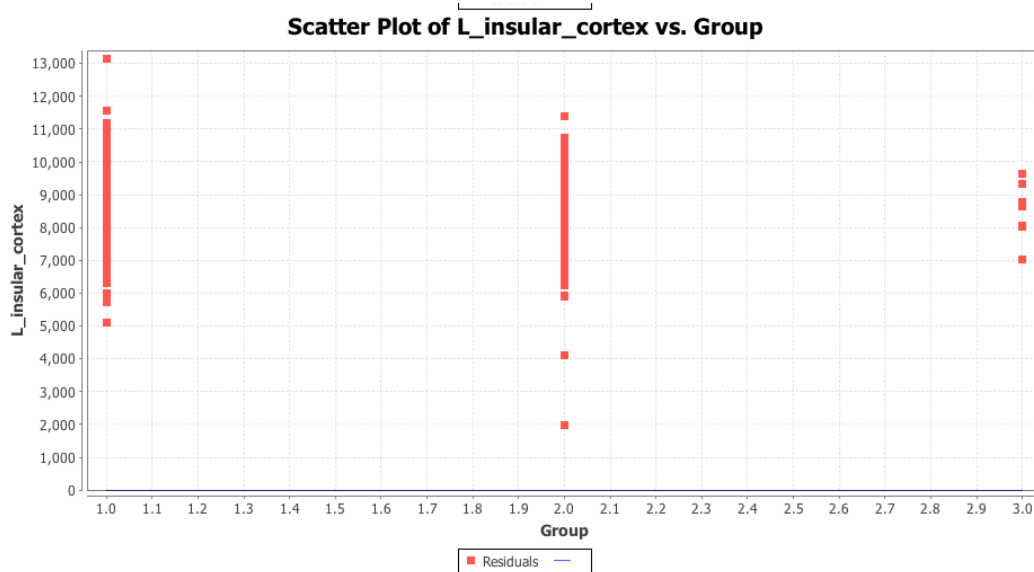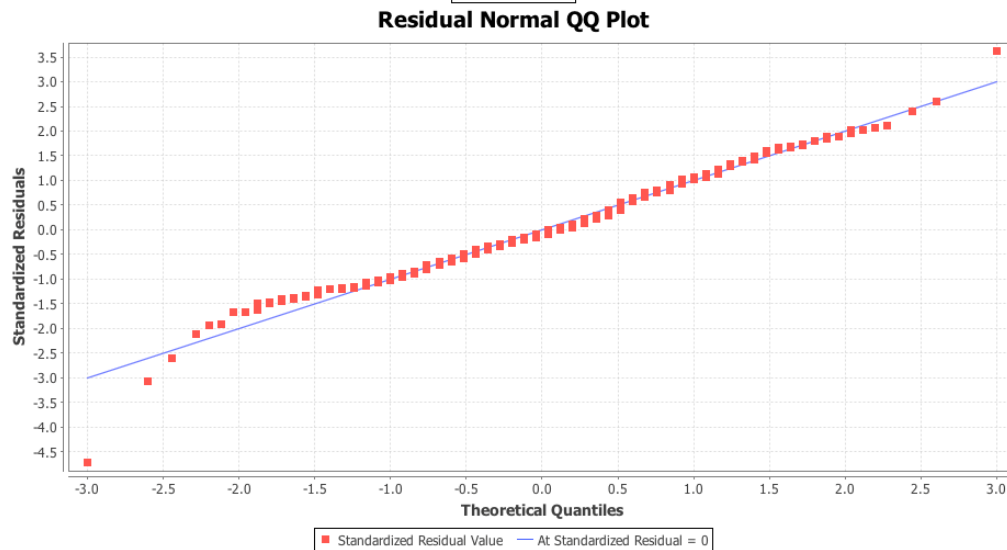
And the numerical results:

Sample Size = 335
Dependent Variable = L_insular_cortex
Independent Variable(s) =   Group   Sex   Interaction Group:Sex

*** Two-Way Analysis of Variance Results ***

Standard 2-Way ANOVA Table. See:
http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_ANOVA_2Way
==
VarianceSource	DF	RSS	MSS	F-Statistics	P-value
MainEffect:Group	2	12389469.895	6194734.948	3.695	.026
MainEffect:Sex	1	230366.224	230366.224	.137	.711
Interaction Group:Sex	2	2086365.301	1043182.650	0.58	0.57
Error	329	551643331.746	1676727.452		
Total:	334	566738725.415			
==

Variable: Group

Degrees of Freedom = 2
Residual Sum of Squares = 12389469.895
Mean Square Error = 6194734.948
F-Value = 3.695
P-Value = .026

Variable: Sex
Degrees of Freedom = 1
Residual Sum of Squares = 230366.224
Mean Square Error = 230366.224
F-Value = .137
P-Value = .711

Variable: Interaction Group:Sex
Degrees of Freedom = 2
Residual Sum of Squares = 2086365.301
Mean Square Error = 1043182.650
F-Value = 0.58
P-Value = 0.57

Residual: Degrees of Freedom = 329
Residual Sum of Squares = 551643331.746
Mean Square Error = 1676727.452
F-Value = 1.801
P-Value = 0.11220599895301875

R-Square = .022

Problem 3: Determine a [linear model](#) of association between the response (Total Brain Volume, TBV) and the following predictors (Group, Sex, and Age). Explain your findings and interpret your model results.

Here is a graphical representation of the data:
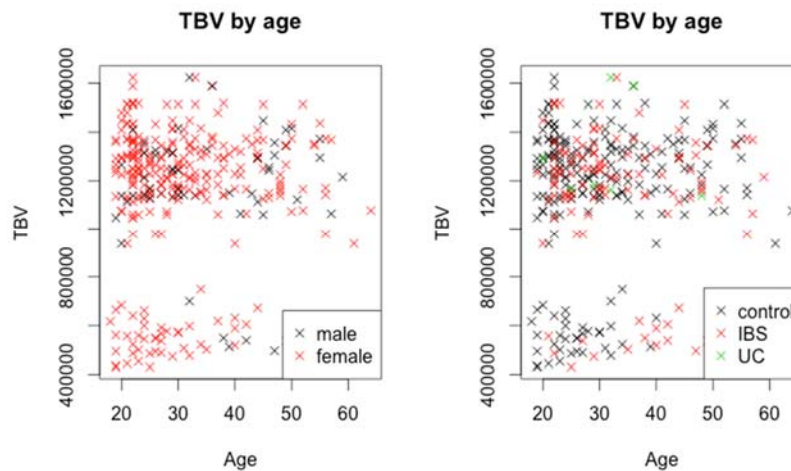
There does not appear to be a strong linear relationship between the two continuous variables (i.e., TBV and Age) nor do any patterns between males and females immediately strike me. I do notice that there are two clusters in the data, one of high total brain volumes and one of low. When looking at the second panel, in which the points are colored according to their group affiliation, we can note that there are no green points (associated with ulcerative colitis) in the bottom cluster. However, there are also very few green points in general, so this pattern might occur frequently by chance. We will now explore these observations quantitatively using a linear model.

```
> fit2.tbv<-lm(TBV~Group + Age + Sex, data=neuro.dat)
> summary(fit2.tbv)
```

Call:
lm(formula = TBV ~ Group + Age + Sex, data = neuro.dat)

Residuals:
    Min   1Q   Median   3Q   Max
  -709562 -52621 77056   170578 488564

Coefficients:

|	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1104030	67894	16.261	<2e-16 ***
Group2	-1758	33866	-0.052	0.959
Group3	149347	109664	1.362	0.174
Age	1908	1550	1.231	0.219
Sex2	-10384	43301	-0.240	0.811

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 283600 on 330 degrees of freedom
Multiple R-squared: 0.0112,   Adjusted R-squared: -0.0007876
F-statistic: 0.9343 on 4 and 330 DF, p-value: 0.4442

The only coefficient with a significant p-value is the intercept. This means that the intercept for the linear model is significantly different from 0 – a result that is not particularly surprising or interesting. None of the independent variables significantly predict TBV (e.g., all p-values are greater than 0.1). Additionally, looking at the coefficient of determination, R-squared, we can see that very little of the variation in the data is explained by the model (i.e., the number is quite close to 0). In fact, the adjusted R-squared, which takes into account the fact that additional parameters may not actually be providing useful additional information, is negative.

Here is the parallel analysis and its results in SOCR:

A useful scatter plot (one of several):
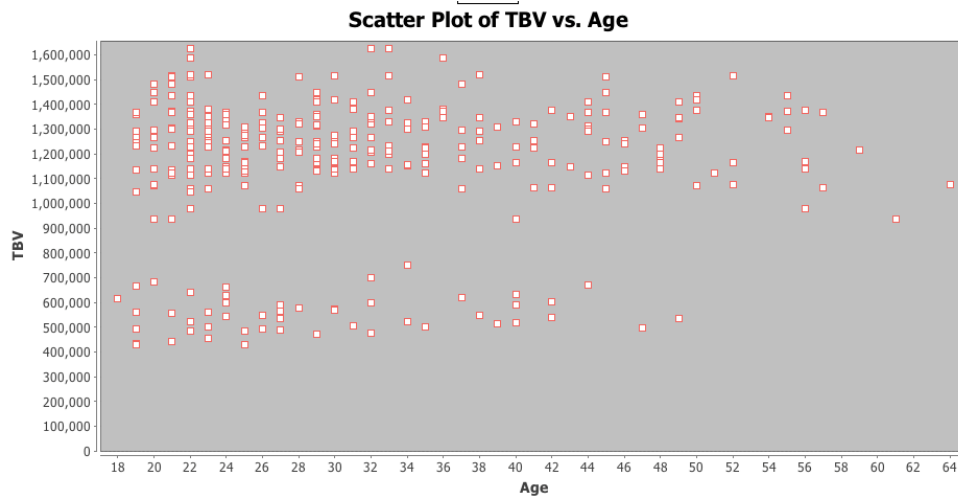
And the numerical results:

Number of Independent Variable(s) = 3
Sample Size =335
Dependent Variable = TBV
Independent Variable(s) =   Sex Group Age

Regression Model:
        TBV = 1101253.901 +1794.449*Age +17419.881*Group -12801.616*Sex + E.

INTERCEPT:
Estimate = 1101253.901
Standard Error = 111651.598
T-Value = 9.863
P-Value = .000

Age:
Estimate = 1794.449
Standard Error = 1548.471
T-Value = 1.159
P-Value = .247

Group:
Estimate = 17419.881
Standard Error = 30223.589
T-Value = .576
P-Value = .565

Sex:
Estimate = -12801.616
Standard Error = 43294.901
T-Value = -.296
P-Value = .768

R-Square = .007

Problem 4: Do people with larger brains necessarily have bigger hippocampal volumes (e.g., L_hippocampus + R_hippocampus)? Formulate a research hypothesis, propose an appropriate statistical technique, apply the corresponding software tool, and report your findings.

H_0: Larger brains are not associated with larger hippocampal volumes.
H_A: Hippocampal volumes are associated with total brain sizes.

To assess this hypothesis, we can generate a linear model of the form:

$$Brain\ size = intercept + \beta * total\ hippocampal\ volume + \varepsilon$$

We can then assess whether the slope of the best fit line is significantly different from 0. Assume again $\alpha=0.05$.

Graphically, we can represent the data and model as follows.
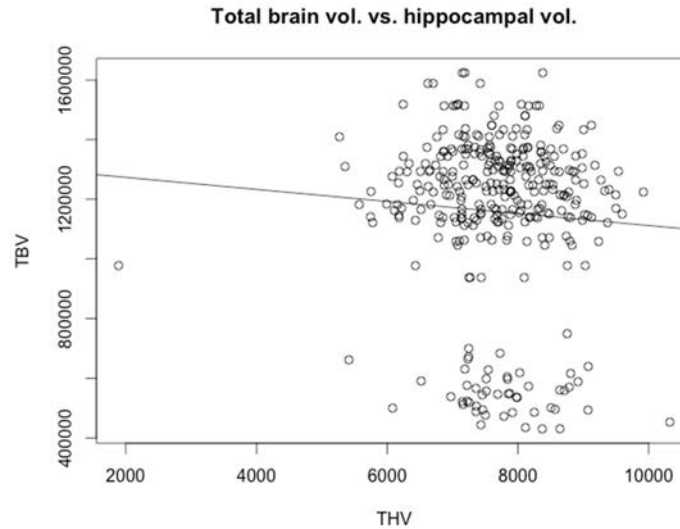
We note a few things about the data. (1) There appears to be a THV outlier. (2) There are two clusters of points with respect to TBV.

> fit.tbm.thv <- lm(TBV~THV, data=neuro.dat)
> summary(fit.tbm.thv)

Call:
lm(formula = TBV ~ THV, data = neuro.dat)

Residuals:
  Min   1Q   Median   3Q   Max
-719034 -46161 75305   176718 481169

Coefficients:
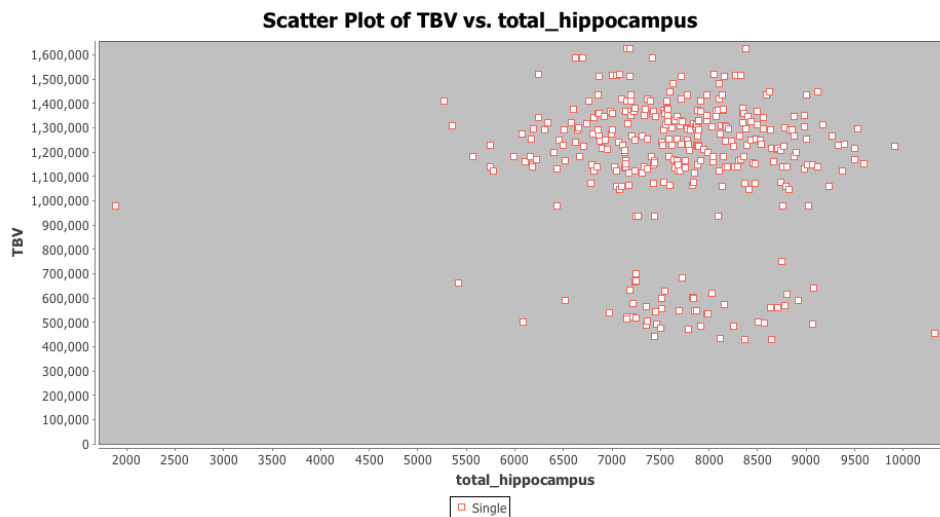|	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1319948.47	132670.25	9.949	<2e-16 ***
THV	-21.12	17.15	-1.232	0.219

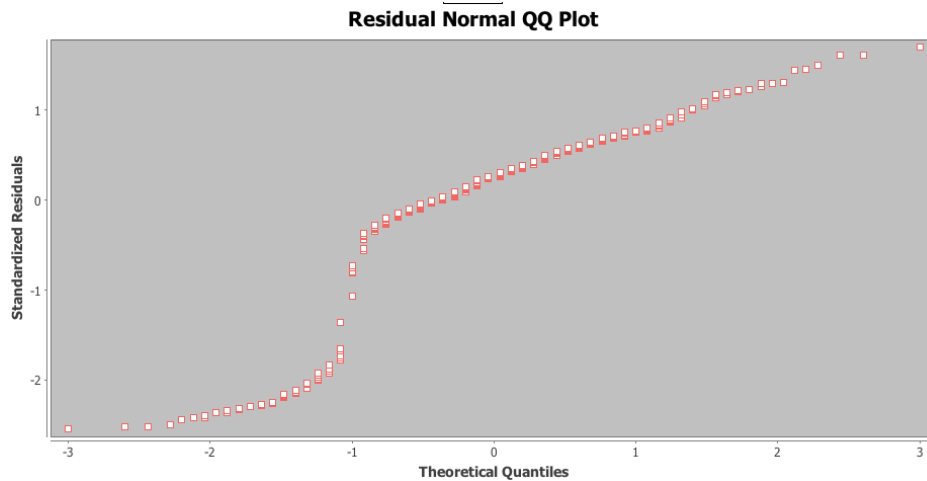Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 283300 on 333 degrees of freedom
Multiple R-squared: 0.004535,   Adjusted R-squared: 0.001546
F-statistic: 1.517 on 1 and 333 DF, p-value: 0.2189

The results suggest that the slope is not significantly different from 0. We therefore fail to reject our null hypothesis and conclude that there is not sufficient evidence to suggest that a relationship exists between hippocampal volumes and brain size. The sign of the coefficient tells us that there is a slight negative relationship between THV and TBV. In particular, an increase of 1 unit of THV is associated with a decrease of 20 units in TBV. However, this relationship is not significant. Additionally, we note the very small R-squared value, indicating that the model does not explain much of the variation in the data.

Here is the parallel analysis and its results from SOCR:

Residual Normal QQ Plot

The lack of a match between the normal distribution and the distribution of the residuals (i.e., the points do not fall on the 1-to-1 line but are instead S-shaped) suggests that perhaps this model is not the most appropriate. For now, however, we will continue with the analysis as is. In the future, we will consider other types of linear model formulations that might better capture the variation in the data.

Here are the numerical results:

      Number of Independent Variable(s) = 1
      Sample Size =335
      Dependent Variable = TBV
      Independent Variable(s) = total_hippocampus

      Regression Model:
          TBV = 1319948.469 -21.119*total_hippocampus + E.

      INTERCEPT:
      Estimate = 1319948.469
      Standard Error = 132670.254
      T-Value = 9.949
      P-Value = .000

      total_hippocampus:
      Estimate = -21.119
      Standard Error = 17.146
      T-Value = -1.232
      P-Value = .219

      R-Square = .005

Appendix: R Code

```
####################
# SMHS851 HW 3
# 10.1.14
# Jennie Lavine
####################

setwd('~/hw3_851')

neuro.dat<-read.csv('neuroimage.csv',header=T)

#######################
#clean data of interest
#######################
neuro.dat$Group<-as.factor(neuro.dat$Group)

table(neuro.dat$Group)
neuro.dat<-subset(neuro.dat, neuro.dat$Group!=5)
#Remove unused levels from the factor 'Group'
neuro.dat$Group<-factor(neuro.dat$Group)

############################
#Take a look at your data
############################
#Side by side boxplots
boxplot(GMV~Group, data=neuro.dat, xlab='Group', ylab='GMV',
    main='GMV by Group',
    sub='Group 1=Control; Group 2=IBS; Group 3=Ulcerative Colitis'
    )

#Histograms of each group
par(mfrow=c(3,1))
hist(neuro.dat$GMV[neuro.dat$Group=='1'])
hist(neuro.dat$GMV[neuro.dat$Group=='2'])
hist(neuro.dat$GMV[neuro.dat$Group=='3'])

#Run an ANOVA
anova.gmv<-aov(GMV~Group, data=neuro.dat)
summary(anova.gmv)

#Fit a linear model, then show ANOVA results (same results as above)
fit.gmv<-lm(GMV~Group, data=neuro.dat)
anova(fit.gmv)

#######################
#Problem 2
#######################
neuro.dat$Sex<-factor(neuro.dat$Sex)
anova.r.ins<-aov(R_insular_cortex~Group*Sex, data=neuro.dat)
summary(anova.r.ins)

anova.l.ins<-aov(L_insular_cortex~Group*Sex, data=neuro.dat)
summary(anova.l.ins)

#########################
#Problem 3
```

```
########################

#This is a full model with all possible interactions
fit.tbv<-lm(TBV~Group*Sex*Age, data=neuro.dat)

par(mfrow=c(1,2))
plot(TBV~Age, data=neuro.dat, col=neuro.dat$Sex, pch=4, main='TBV by age')
legend('bottomright', legend=c('male','female'), pch=4, col=1:2)
plot(TBV~Age, data=neuro.dat, col=neuro.dat$Group, pch=4, main='TBV by age')
legend('bottomright', legend=c('control','IBS', 'UC'), pch=4, col=1:3)

summary(fit.tbv)
anova(fit.tbv)

#This is a model that considers the effect of each variable independently
#It does not take interactions into account
fit2.tbv<-lm(TBV~Group + Age + Sex, data=neuro.dat)
summary(fit2.tbv)




############################
#Problem 4
########################
neuro.dat$THV<-neuro.dat$L_hippocampus + neuro.dat$R_hippocampus
par(mfrow=c(1,1))
fit.tbm.thv <- lm(TBV~THV, data=neuro.dat)
summary(fit.tbm.thv)

plot(TBV~THV, data=neuro.dat, main='Total brain vol. vs. hippocampal vol.')
abline(fit.tbm.thv)




#####################
#Extra code: Calculating summary statistics and plotting for 2-way ANOVA
#####################
sum.stats<-matrix(NA, ncol=6, nrow=11)
colnames(sum.stats)<-c('mean','sd','n','SEM', 'ci90.lo','ci90.hi')
rownames(sum.stats)<-c('group1','group2','group3','male','female',
            'group1:male','group2:male','group3:male',
            'group1:female','group2:female','group3:female')


for(i in 1:3){
  sum.stats[i,'mean']<-mean(neuro.dat$R_insular_cortex[neuro.dat$Group==i])
  sum.stats[i,'sd']<-sd(neuro.dat$R_insular_cortex[neuro.dat$Group==i])
  sum.stats[i,'n']<-sum(neuro.dat$Group==i)
  sum.stats[i,'SEM']<-sum.stats[i,'sd']/sqrt(sum.stats[i,'n'])
}

for(i in 1:2){
  sum.stats[3+i,'mean']<-mean(neuro.dat$R_insular_cortex[neuro.dat$Sex==i])
  sum.stats[3+i,'sd']<-sd(neuro.dat$R_insular_cortex[neuro.dat$Sex==i])
  sum.stats[3+i,'n']<-sum(neuro.dat$Sex==i)
  sum.stats[3+i,'SEM']<-sum.stats[3+i,'sd']/sqrt(sum.stats[3+i,'n'])
}
```

```
for(i in 1:3){
  sum.stats[5+i,'mean']<-mean(
    neuro.dat$R_insular_cortex[neuro.dat$Group==i & neuro.dat$Sex==1])
  sum.stats[5+i,'sd']<-sd(
    neuro.dat$R_insular_cortex[neuro.dat$Group==i& neuro.dat$Sex==1])
  sum.stats[5+i,'n']<-sum(neuro.dat$Group==i & neuro.dat$Sex==1)
  sum.stats[5+i,'SEM']<-sum.stats[5+i,'sd']/sqrt(sum.stats[5+i,'n'])
}

for(i in 1:3){
  sum.stats[8+i,'mean']<-mean(
    neuro.dat$R_insular_cortex[neuro.dat$Group==i & neuro.dat$Sex==2])
  sum.stats[8+i,'sd']<-sd(
    neuro.dat$R_insular_cortex[neuro.dat$Group==i& neuro.dat$Sex==2])
  sum.stats[8+i,'n']<-sum(neuro.dat$Group==i & neuro.dat$Sex==2)
  sum.stats[8+i,'SEM']<-sum.stats[8+i,'sd']/sqrt(sum.stats[8+i,'n'])
}

sum.stats[,'ci90.lo']<-sum.stats[,'mean']-1.64*sum.stats[,'SEM']
sum.stats[,'ci90.hi']<-sum.stats[,'mean']+1.64*sum.stats[,'SEM']


#Plot effect of group
par(mfrow=c(1,2))
plot(as.factor(1:3),sum.stats[1:3,'mean'], ylim=c(4000,7100),
    xlab='Group', ylab='R_insular_cortex', main='Effect of Group', type='n')
points(sum.stats[,'ci90.lo'], col=2)
points(sum.stats[,'ci90.hi'], col=2)

plot(as.factor(1:2),sum.stats[4:5,'mean'], ylim=c(5000,6000),
    xlab='Sex', ylab='R_insular_cortex', main='Effect of Sex')
points(sum.stats[4:5,'mean']+1.96*sum.stats[4:5,'SEM'], col=2)
points(sum.stats[4:5,'mean']-1.96*sum.stats[4:5,'SEM'], col=2)

par(mfrow=c(1,1))
plot(1:3,sum.stats[6:8,'mean'], ylim=c(4500,6600),
    xlab='Group', ylab='R_insular_cortex', main='Effect of Group and Sex \n with 90% CI', type='b')
points(sum.stats[6:8,'mean']+1.64*sum.stats[6:8,'SEM'], col=1, pch=6)
points(sum.stats[6:8,'mean']-1.64*sum.stats[6:8,'SEM'], col=1, pch=2)
lines(1:3, sum.stats[6:8,'mean'], lwd=3)

points(as.factor(1:3),sum.stats[9:11,'mean'],col=3)
points(sum.stats[9:11,'mean']+1.64*sum.stats[9:11,'SEM'], col=3, pch=6)
points(sum.stats[9:11,'mean']-1.64*sum.stats[9:11,'SEM'], col=3, pch=2)
lines(1:3, sum.stats[9:11,'mean'], lwd=3,col=3)

legend('top',legend=c('male','female'),fill=c(1,3))
```